Developing Metadata Application Profiles

Mariana Curado Malta
Polytechnic of Oporto, Portugal & Algoritmi Center, University of Minho, Portugal

Ana Alice Baptista
Algoritmi Center, University of Minho, Portugal

Paul Walk
University of Edinburgh, UK

A volume in the Advances in Web Technologies and Engineering (AWTE) Book Series

www.igi-global.com

Published in the United States of America by
IGI Global
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Malta, Mariana Curado, 1969- editor. | Baptista, Ana Alice, editor. | Walk, Paul, 1968- editor.
Title: Developing metadata application profiles / Mariana Curado Malta, Ana Alice Baptista, and Paul Walk, editors.
Description: Hershey, PA : Information Science Reference, [2017] | Includes bibliographical references and index.
Identifiers: LCCN 2016056939| ISBN 9781522522218 (hardcover) | ISBN 9781522522225 (ebook)
Subjects: LCSH: Metadata--Standards.
Classification: LCC Z666.7 .D48 2017 | DDC 025.3--dc23 LC record available at https://lccn.loc.gov/2016056939

This book is published in the IGI Global book series Advances in Web Technologies and Engineering (AWTE) (ISSN: 2328-2762; eISSN: 2328-2754)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 3
The Minimum Mandatory Metadata Sets for the KIM Project and RAIDmap

Alexander Ball
University of Bath, UK

Mansur Darlington
University of Bath, UK

Christopher McMahon
University of Bristol, UK

ABSTRACT

A Minimum Mandatory Metadata Set (M3S) was devised for the KIM (Knowledge and Information Management Through Life) Project to address two challenges. The first was to ensure the project's documents were sufficiently self-documented to allow them to be preserved in the long term. The second was to trial the M3S and supporting templates and tools as a possible approach that might be used by the aerospace, defence and construction industries. A different M3S was devised along similar principles by a later project called REDm-MED (Research Data Management for Mechanical Engineering Departments). The aim this time was to help specify a tool for documenting research data records and the associations between them, in support of both preservation and discovery. In both cases the emphasis was on collecting a minimal set of metadata at the time of object creation, on the understanding that later processes would be able to expand the set into a full metadata record.

DOI: 10.4018/978-1-5225-2221-8.ch003

INTRODUCTION

Between 2005 and 2013 the University of Bath was involved in a series of linked projects aimed at improving knowledge and information management in engineering. The first of these went by the title 'Immortal Information and Through-Life Knowledge Management: Strategies and Tools for the Emerging Product–Service Paradigm', though that was colloquially abbreviated to 'Knowledge and Information Management Through Life' and thence to 'KIM'.

The KIM Project was a Grand Challenge project funded by the Engineering and Physical Sciences Research Council (EPSRC) and the Economic and Social Research Council (ESRC) in the UK (Ball et al., 2006). It ran for three-and-a-half years and involved 11 universities and numerous industrial collaborators. One of the purposes of the KIM Project was to experiment with new ways of working on engineering projects in order to increase efficiency across the lifecycle, and to improve the long-term usability of the project records. This being the case, the project incorporated some of those ways of working into its own governance and processes as both a test and a demonstrator. For example, all project files were given a coded file name that indicated the work unit to which it belonged, the type of document, the initial creator, and the version, but did not reveal the content. Instead, researchers were expected to embed the title in the document properties, and a separate registry was maintained that decoded file names into document titles.

One of the methods used to protect the longevity of the project records was to impose a Minimum Mandatory Metadata Set (M3S) for all project documents. Researchers were required to embed the specified metadata in the documents they created. The intention was to use it, alongside regular file properties and information supplied at the collection level, to generate a complete set of preservation and descriptive metadata for each document. In order to reduce the burden this would place on researchers, document templates were written that used the embedded metadata to fill out content on title pages, headers, footers, and so on. As mentioned above, additional metadata was encoded in the file name convention.

Even though the focus of the KIM Project was on knowledge and information management within industry, some aspects of the work had wider applicability. Therefore, when the Joint Information Systems Committee (JISC) of the UK further and higher education funding councils set up a programme to fund projects tackling various challenges in the area of research data management (RDM), the University of Bath took forward some of the ideas from KIM and applied them to academic research in the course of two much smaller projects.

ERIM (Engineering Research Information Management) developed a set of RDM processes for the Innovative Design and Manufacturing Research Centre at Bath, including a technique for visualizing the inter-relationships between the various

records generated by a research project. The latter technique, known as Research Activity Information Development (RAID) modelling, was based on UML activity diagrams and could be used to trace the results published in a paper back to raw data files or further back to project plans (Ball, Darlington, Howard, McMahon, & Culley, 2012).

The follow-on project REDm-MED (Research Data Management for Mechanical Engineering Departments) generalized these processes for use by the Mechanical Engineering Department – in consultation with a parallel project setting up RDM support across the university – and by engineering departments in other universities. The project also developed a software tool called RAIDmap to aid in the creation of RAID diagrams.

In the event the digital RAID diagrams produced by RAIDmap were much richer than the purely graphical diagrams developed by ERIM. They were capable of storing metadata records for each node (file, instrument) in the diagram, populated partly automatically and partly manually. Indeed, the vision for RAIDmap was that it would monitor the user's workspace and automatically add nodes and metadata to the diagram as the user worked; the user would periodically fill in any gaps in the metadata and note relationships between nodes. The question naturally arose of what metadata users should be asked to provide themselves, and what the best practice should be for providing it: embedding it within the document or entering it manually into RAIDmap. To that end, the idea of a Minimum Mandatory Metadata Set was revisited.

This chapter presents the two Minimum Mandatory Metadata Sets and the process by which they were developed and implemented. It explores the rationale and motivation for the respective sets, and reflects on the merits of the approach taken.

BACKGROUND

In discussing the metadata requirements of the various stakeholders involved, the authors found it helpful to use the terminology defined by the Open Archival Information System (OAIS) Reference Model (Consultative Committee for Space Data Systems, 2002). The Information Model presented in that standard defines five types of metadata that have an important role when preserving data objects in an archive:

- **Representation Information:** The information needed by the user to interpret and understand the data object. An archive would be expected to store enough Representation Information to satisfy the needs of a typical member of the Designated Community, that is, the user group that the archive has

committed to support. The collective name for the data object and its associated Representation Information is Content Information.

- **Provenance Information:** Information about the source of the Content Information, the chain of custody since its creation and the operations performed on it.
- **Context Information:** Information describing how the Content Information relates to other information resources.
- **Reference Information:** Unique identifiers for the Content Information.
- **Fixity Information:** Checksums or other information that could be used to detect, and possibly reverse, undocumented alterations to the Content Information.

A later version of the standard defines a further class of metadata, Access Rights, consisting of the permissions, licenses and terms of use for accessing, preserving, distributing, and using the Content Information (Consultative Committee for Space Data Systems, 2012). When a data object is packaged together with the above types of metadata, it forms what OAIS calls an Information Package.

While these terms were useful for describing the various metadata that might be of relevance, the OAIS Information Model did not go so far as to enumerate the various metadata elements that would be needed in each category. The authors therefore looked at other work that had been done in the area of preservation metadata.

To serve the needs of its digital collections, the National Library of Australia (1999) proposed a scheme of 25 elements, of which two were complex (i.e., had subelements). The starting point for the proposal was a 16-element set recommended by RLG for images generated by digitization projects (Research Libraries Group, Working Group on Preservation Issues of Metadata, 1998). The NLA version was at once more detailed and general: it could be applied to audiovisual material, databases, executables, and text as well as images. The OAIS model was one of the influences for the scheme, as was the data model for PANDORA, Australia's Web archive, along with metadata schemes being developed concurrently by other groups.

One of these was the scheme developed by the Cedars project (CURL Exemplars in Digital Archives, 2000) for university research libraries seeking to preserve archival digital content. This scheme was explicitly based on the OAIS Information Model: it was conceived as a hierarchy of complex elements, with the upper levels reproducing the OAIS model of an Information Package, and the lower levels filling in the details under the above categories. To give an impression of the size of the scheme, the hierarchy terminated in 24 simple elements. Notably, it introduced a detailed tree of elements for rights management that it included under Provenance Information.

Another was developed in the context of the Networked European Deposit Library (NEDLIB) project. Lupovici & Masanès (2000) proposed what they described as a minimal metadata set supporting the preservation of digital publications, aimed at national libraries. It consisted of eight top-level elements: five for Representation Information – one for each of the five layers in the OAIS Layered Information Model – and one each for Reference, Provenance and Fixity Information. The scheme could be said to be minimal in that it excluded rights and Context Information as understood above, but was actually more detailed than the Cedars scheme. The top-level elements were structured with 38 sub-elements, only six of which were themselves grouping further sub-elements.

In response to this flurry of activity, OCLC and RLG convened a Working Group on Preservation Metadata to build a consensus on best practice. They produced two reports in quick succession. The first was a white paper that compared and mapped between the above three schemes, and drew additional insights from an XML schema used by Harvard University's Data Repository Services (OCLC/RLG Working Group on Preservation Metadata, 2001). The second went a step further, using that comparison to merge the NLA, Cedars, and NEDLIB schemes into a unified hierarchy of elements (OCLC/RLG Working Group on Preservation Metadata, 2002). Some elements were added or refined after suggestions by Working Group members, and the elements relating to Fixity Information were taken instead from a scheme used by OCLC for its Digital Archive product. This hierarchy was called a framework rather than a scheme as it was not fully fleshed out: it did not specify how the information should be recorded for each element, for example, but suggested that some elements may need to be decomposed into (further) sub-elements. Also, it did not give much assistance on dealing with complex data objects, where an element from the framework may need a different value depending on which component file was considered.

The OCLC/RLG framework was intended as a foundation for future work rather than an end in itself. When the National Library of New Zealand (NLNZ) came to devise a scheme for its digital holdings, it concurred with the principles behind it, but felt that the NLA proposal was a better starting point for a practical and implementable metadata scheme (National Library of New Zealand, 2003). While the framework had favoured the OAIS Information Model for its basic structure, the NLNZ scheme returned to the NLA's data model of collections, objects, and files, and a more linear sequence of elements. Indeed, it went further and introduced a new four-entity model where elements could apply to objects (simple, complex, or groups), files, processes, or metadata modifications.

OCLC and RLG convened a second working group in 2003 to tackle much the same issue: how to turn the earlier framework into something practical and implementable. The group, Preservation Metadata: Implementation Strategies (PREMIS),

came to very similar conclusions to the NLNZ scheme. It developed a five-entity data model of Intellectual Entities (similar to the NLA's idea of a 'work'), Objects (similar to the NLA's idea of a 'manifestation'), Events (similar to the NLNZ's idea of a process), Agents, and Rights. As with the NLNZ scheme, PREMIS recognized three types of object: a file, a bitstream (such as the audio and video streams within a multimedia file), and a representation (a set of files used at once to make a single rendition). Metadata elements were associated directly with these entities instead of being grouped in the OAIS metadata categories. Even more so than earlier schemes, the PREMIS scheme used a hierarchy of sub-elements to allow information to be structured for the benefit of automated tools (OCLC/RLG Preservation Metadata: Implementation Strategies Working Group, 2005).

The PREMIS scheme was warmly welcomed by the digital curation community, and indeed an active user community grew up around it, most conspicuously in the form of the PREMIS Implementors' Group. The Library of Congress agreed to host a Maintenance Activity to support PREMIS and develop it further; it published version 2.0 of the Data Dictionary in 2008 and version 3.0 in 2015. Since the PREMIS activity is focused on implementation, it has produced practical resources including an XML schema, an OWL ontology and a directory of examples of real-world usage (Library of Congress, 2016).

The PREMIS Data Dictionary was in this way the culmination of activity around preservation metadata from a digital repository perspective. There had however been a parallel set of developments among traditional archives (Caplan, 2006; Day, 2004). One strand began with the Functional Requirements for Evidence in Recordkeeping project (University of Pittsburgh, School of Information Sciences, 1996), also known as the Pittsburgh Project. This defined a scheme of elements intended to preserve the evidential value of electronic records over the long term; hence it included concepts, such as the transaction of which the object is a record, that were absent from the repository-based schemes. The scheme had six 'layers' grouping a total of 17 top-level elements and 48 sub-elements, only one of which grouped further sub-elements. About half of the elements and sub-elements were in the structural layer, and related to how the record might be rendered in a computing environment.

This work inspired several other metadata initiatives and was particularly influential in Australia. The *Recordkeeping Metadata Standard for Commonwealth Agencies* (National Archives of Australia, 1999) defined information that the agencies should collect about their records. This scheme was larger in terms of number of elements than the Pittsburgh scheme, but this was due to its having a broader scope. It was shallower in detail, with only four sub-elements dedicated to structural information compared to the 27 in the Pittsburgh scheme, for example. The scheme was fully revised in 2008 and renamed the Australian Government Recordkeeping Metadata Standard; further revisions were made in 2011 and 2015 (National Archives of Aus-

tralia, 2015). Like the NLNZ scheme and PREMIS, it was reconstructed around a multi-entity model – in this case the five entities of Record, Agent, Business, Mandate and Relationship – with a different subset of the 26 elements applying to each entity.

The Victorian Electronic Records Strategy (Public Record Office Victoria, 2000) defined a self-documenting exchange format for records to permit them to be transferred reliably between systems. The metadata scheme for this format essentially encapsulated the Recordkeeping Metadata Standard with additional fields to aid transfer, such as descriptive metadata and encoding information.

Another strand of work began with the Preservation of the Integrity of Electronic Records project (Duranti, Eastwood, & MacNeil, 1997), which among other things defined a set of metadata attributes for records and dossiers. The emphasis of the project was more on procedures and protocols for the safe and auditable handling of electronic records, rather than on the technical challenges of preserving them, thus preservation metadata does not especially feature in the attribute list. The work was however carried forward by the International Research on Permanent Authentic Records in Electronic Systems project ("InterPARES project," n. d.), which began in 1998. The first phase of InterPARES reviewed the whole notion of what preserving electronic records entails; the second, between 2002 and 2007, performed a survey of relevant metadata standards; the third produced learning materials regarding preservation metadata and an application profile for authenticity metadata (Rogers & Tennis, 2016). As the name suggests, the profile was focused on authenticity and evidential integrity rather than preservation as such, with 10 sub-elements dedicated to structural metadata and upwards of 40 dedicated to provenance and context.

THE KIM MINIMUM MANDATORY METADATA SET

The motivation behind the Minimum Mandatory Metadata Set for the KIM Project was to address a particular issue affecting the construction industry and certain engineering industries such as aerospace and defence. These industries supply products that are expected to have a service life of multiple decades: products such as aircraft carriers or hospitals. The phenomenon addressed by the KIM Project was a movement in these industries towards a product–service paradigm, whereby instead of simply selling the product to the customer, the contractor would lease the implied capability to them. The contractor would therefore retain both the ownership of the product and the responsibility for ensuring it continued to perform to the customer's satisfaction.

Under this regime it is more important than ever that records produced early in the design stage of the product remain accessible and, crucially, reusable all the way through to the eventual disposal of the product. Such records may prove vital for

performing maintenance, diagnosing unexpected behaviour, adapting the product to evolving customer needs, and disposing safely of the product once it has reached the end of its service life. The challenge faced by industry was, and is, how to curate digital records over time periods of five decades or more when the organization is likely to change key components of its software environment at intervals of between three and ten years.

Among the proposals for addressing the issue was that companies should systematically collect metadata for the digital records they create: metadata that would assist a digital archivist in preserving the records in the face of technological change. The question that arose naturally from this proposal was what these metadata might be. A literature review conducted in mid-2006 revealed the resources described above, specifically PREMIS version 1.0, the Recordkeeping Metadata Standard for Commonwealth Agencies, and the Victorian Electronic Records Strategy Metadata Scheme. While the first phase of the InterPARES project had concluded by that point it had not produced a workable metadata scheme; that would be a focus of later phases.

The three candidate schemes had a significant degree of overlap, with differences being more a matter of emphasis than of substance. For example, all three recorded information about the digital object, and the chain of processes it has undergone, but PREMIS emphasized the former while the recordkeeping standards emphasized the latter. Similarly, all three provided the means to prioritize the preservation of certain aspects of the digital object. The recordkeeping standards expressed this in terms of a mandate – legal or other requirements that stipulate what must be kept and for how long, or to put it another way, the evidential role or roles that the record would be expected to play. PREMIS used the more abstract and flexible notion of significant properties. The choice was made somewhat harder in that the recordkeeping standards fitted well with the corporate philosophy and environment of the industries with which the project was working, but the interest of the project team was in tackling the preservation challenges of some particularly difficult types of data, for which the PREMIS approach seemed better suited.

The decision was postponed while the next logical question was addressed: how the metadata were to be collected. The project's industrial collaborators were on the whole larger organizations, where there would be a division of labour between those creating the digital objects and those archiving and preserving them. Many had set up dedicated records management systems. A digital archivist in this context would be able to collect a certain amount of information from analysing files as they were ingested into the system, and the system would generate metadata itself as it performed ingest, archiving, and preservation processes on the files. Some information, though, could not be taken for granted: the originator of the file would either

have to provide it separately or embed it deliberately within the file for the archivist to extract. This suggested to the project team that a more interesting question to ask was this: what could the archivist not be expected to deduce regardless of the nature of the digital object, and what information, therefore, would the originator be expected to provide in order to plug that gap. There was a tension to address between completeness and effort: if the team recommended that companies require originators to provide too much additional information, the impact on staff workload would make it unlikely that the recommendation would be adopted; or if it were adopted, it may result in staff ignoring or subverting the requirement.

In order to answer the question and test the result, the project team decided to apply the same principles to the digital objects produced within the project itself, and introduce a minimum mandatory metadata set. This would be a set of metadata that each researcher would be required to record for each project document. It would not be a precise analogy for introducing such a requirement in an industrial setting, but there were enough parallels between the two situations to make it an interesting trial. There was again a separation of roles between the researchers originating the digital objects and the repository managers and digital archivists who would eventually preserve them. Within each university, there would be a degree of consistency in the software used by the project team, as one would also expect in a corporate setting. The majority of researchers on the project were looking at issues other than the one in hand, so their tolerance for additional workload would not be extended by an ideological commitment to metadata.

The authors' approach was to derive the Minimum Mandatory Metadata Set (M3S) by taking a full preservation metadata scheme and subtracting the elements that a repository or archive would be able to supply from automated analysis or internal policy. For the remaining elements, they would suggest how authors could provide the information in the most efficient way, so as to reduce the additional effort. Since this would focus on the properties of the documents, rather than processes, it was decided to use PREMIS as the basis for the M3S; it was also felt that this would be more familiar to the repository managers and archivists who would be most likely to receive the documents at the end of the project.

Derivation of the Set

The first stage of the derivation was to work through the top-level metadata elements for each of the entities in the PREMIS data model. Note that in the following discussion, some levels of detail have been omitted for brevity; also, since the M3S relates to properties of a document, the more trivial translations of these properties into relationships with entities have been omitted.

The following elements are associated with *objects*:

- **Object Identifier:** Archives normally assign these for their own purposes, but for the sake of coordination and consistency across project partners (and indeed archives) it was felt important that the project should also provide its own identifiers, to be applied by the document originators.
- **Preservation Level:** This refers to whether the archive will provide bit-level preservation only, or will use techniques such as format migration or emulation to avoid rendering problems as formats become obsolete. This was for archival policy, rather than document originators, to decide.
- **Object Category:** This refers to whether the object in question is a representation (group of files), a single file, or a bitstream within a file. On a practical level, the authors did not expect document originators to provide metadata on a bitstream level, and anticipated that most documents would be represented by a single file.
- **Object Characteristics:** These were considered at the sub-element level:
  - **Composition Level:** This refers to the number of encryption or compression operations that have been applied to an object; it could be supplied implicitly by requiring originators to submit their documents uncompressed and unencrypted.
  - **Fixity:** Checksums could be generated by the archive.
  - **Size:** This could be measured by the archive.
  - **Format:** This could be derived from (a) the filename extension, (b) knowledge of the software environment of the project team members, and (c) information embedded within files by software automatically.
  - **Significant Properties:** This was again a matter of archive policy.
  - **Inhibitors:** This refers to password protection and similar mechanisms, which the project did in fact employ. For obvious reasons, passwords could not be recorded in the document, so the authors resolved to require originators to supply the passwords separately to the archive on ingest.
- **Creating Application:** As for format, this could be derived from (a) the filename extension, (b) knowledge of the software environment of the project team members, and (c) information embedded within files by software automatically.
- **Original Name:** The filename used by the originator, by definition.
- **Storage:** This refers to the location of the object after ingest.
- **Environment:** The authors resolved to collect information about the software and hardware environment for each university's team and record this at the collection level, instead of asking originators to provide it on a per-document basis.

- **Signature Information:** The project had no plans to use digital signatures; if they were applied at all, it was felt likely to be after ingest.
- **Relationships:** There may be various indirect relationships between project documents (e.g., belonging to the same task) but the authors felt that direct relationships should be specified at the point of ingest rather than embedded at the time of document creation.
- **Linking Intellectual Entity Identifier:** This is an artifact of the data model, allowing multiple objects to instantiate the same work. Such relationships could be implied through a file naming convention; for example, the same work in Microsoft Word and Portable Document Format would have the same filename but a different filename extension for each format, and different versions of a file would have the same name apart from the version number.

PREMIS records the processing history of objects as *events*. The important pre-ingest events from the project perspective were the date of creation, date of issue (i.e., when it was last saved before being submitted for approval), date of approval, and date of last modification (typically to record the fact of approval). The following elements are associated with events:

- **Event Identifier:** This is an artifact of the data model and was left to the archive to assign.
- **Event Type, Event Date Time:** Document originators were best placed to provide this information. The authors felt it would be most intuitive for them to do so in the form of key–value pairs, with the type as the key and the date as the value.
- **Event Detail:** By using a standard set of date types, the project hoped to avoid the need for further clarifying text.
- **Event Outcome Information:** This allows archives to record the outcome of a process such as checking a file against its recorded checksum. It did not seem to be applicable to the above four date types.
- **Linking Agent Identifier:** This relationship includes a role, which in the case of the project could be derived from the date type, i.e. 'authorizer' (or similar) for the date of approval, and 'creator' for the others. It was agreed that the originator would provide the identities of the lead author and the person who approved the document.

The intention was for the lead author and authorizer to be recorded in PREMIS as *agents* with agent type 'person'. The project would supply a list of project members and the codes used internally in the project to identify them (corresponding

to PREMIS elements "agent name" and "agent identifier") to allow the archive to associate them correctly with its own internal identifiers.

Version 1 of PREMIS acknowledged the importance of recording rights information by including a *rights* entity, but did not specify what would be important to record from a preservation perspective. Meanwhile, the project team had implemented an internal information classification scheme as a way of controlling access to commercially sensitive information:

- **Working:** The document is in draft and only for distribution by and discussion with the lead author.
- **Level 1:** The document is unrestricted and may be disseminated openly.
- **Level 2:** Access to the document is restricted to those registered with the KIM Project (in effect, the project team, the industrial collaborators and other interested academics and researchers).
- **Level 3:** Access to the document is restricted to members of the project team only.
- **Level 4:** Access to the document is restricted to a specified distribution list.

A complete expression of this access control in PREMIS would consist of the following elements:

- **Permission Statement Identifier:** This could be left for the archive to assign.
- **Granting Agent:** It was understood that part of the approval process for documents was to authorize dissemination at the stated access level, so the most appropriate agent to identify here would be the authorizer.
- **Granting Agreement:** This could be generated by the archive.
- **Permission Granted:**
    - **Act:** From the suggested controlled vocabulary in the PREMIS Data Dictionary, the most appropriate term would be 'disseminate'.
    - **Restriction:** For documents at Access Levels 2–4, this should indicate that dissemination would be limited to the given list of individuals. It was agreed that for Access Levels 2–3, this list would be provided by the project, while for Access Level 4, the list would be provided by the originator.
    - **Term of Grant:** The intention was to negotiate this at the time of ingest.

The last entity in PREMIS to consider was the *intellectual entity*. PREMIS does not enumerate metadata elements for this entity, but refers the reader to descriptive metadata schemes such as MARC, MODS and Dublin Core. The authors therefore had to consider which elements would be most useful to enable the discovery and

retrieval of project documents. For the project's own purposes, they felt the most useful access points would be as follows:

- **Authors:** The full list of authors, as opposed to just the lead author.
- **Title:** A succinct statement of the subject matter.
- **Keywords:** The major topics of the document.
- **Document Type:** The project had established different series of documents; the most common types were reports and presentations, but there were also agenda, minutes, visit reports, discussion documents, tools (e.g., templates, questionnaires, participant briefings) and internal communications (e.g., procedures, directives, guidance).
- **Task Code:** This could be used as OAIS Context Information, grouping together documents relating to the same strand of research within the project.

After that first pass, the authors were left with four lists of information: that which would be specific to the document, and which the originator should embed within it; that which the originator should provide separately at ingest; that which would be common to all documents, and which the project should provide at the point of ingest; and that which would be generated by the archive. It was the first list that formed the basis of the M3S:

- Document identifier (at the level of intellectual entity and file)
- Filename, including extension and version number
- Title
- Lead author and other authors
- Date created, issued, approved, last modified
- Approved by
- Keywords
- Document type
- Task code
- Access level and, if Level 4, distribution list

The next stage was to rationalize the metadata to enable it to be recorded in the most efficient way. The approach taken was to encode as much information as possible in the identifier, which would also serve as the filename.

In PREMIS terms, the intellectual entity identifier was made up of five elements:

- A code identifying it as a KIM Project document ('kim');
- The task code, represented by two integers (e.g., '12' for Work Package 1, Task 2);

- The document type, represented by a three-letter code (e.g., 'rep' for report);
- A three-digit rank number, with each task code and document type having its own sequence starting at '001' – in order to support this, the project had to set up a central register to allocate the numbers;
- The lead author, represented by a two- or three-letter code.

Filenames were formed by appending two further pieces of information:

- The version number, normally expressed as two integers (e.g. '13' for version 1.3) but in the case of a document edited by someone other than the lead author, extended with a hyphen, the date in YYMMDD format, and the two- or three-letter code for the editor (e.g., '-060312ab');
- The normal filename extension.

In order to make the provision of the remaining information as painless as possible, the project produced a series of templates that made use of embedded metadata or document properties and exposed them as part of the title page information. For example, the Microsoft Word templates used the built-in document properties *Title*, *Author*, and *Keywords*, the custom properties *Access level*, *Approved by*, and *Date approved*, and the built-in metadata for the date and time of creation and last modification. The report template displayed most of these properties on the title page so the author did not have to type them twice. The template also displayed the filename and a generated issue date on the title page, and the title in page headers, using the same mechanism.

Considered thus far, the implementation of the M3S could not be considered a schema or profile. While the conventions in place would allow the metadata to be extracted by an automated script, only those elements taken from built-in document properties could be interpreted by generic tools. The semantics of the remaining elements would have to be inserted by custom scripts.

The opportunity to transform the M3S into an application profile, in the sense defined by Heery and Patel (2000), presented itself when the authors considered how to embed the metadata in PDF files. While the PDF specification did not provide direct support for custom properties, it allowed them to be added indirectly by means of an Extensible Metadata Platform (XMP) packet (Abode Systems, n.d.); XMP packets are XML documents that express properties of the containing document using RDF/XML. The M3S was therefore translated into RDF (see Table 1) so that it could be inserted into PDF files, either at creation time by the LaTeX templates or after the fact using a metadata panel add-on for Adobe Acrobat. Some compromises were made due to a real or perceived lack of entirely suitable predicates (such as

Table 1. KIM M3S expressed using RDF predicates. The namespaces dc, dcterms and owl have their usual meanings; xap refers to Adobe's XMP Basic Schema

Metadata element	RDF predicate	Value
Object identifier	dcterms:isFormatOf	Partial filename
Filename	dc:identifier	Full filename
Task code	xap:Label	'KIM Task *n.m*' or (if $m = 0$) 'KIM Work Package *n*'
Document type	dc:type	'Report, 'Presentation', etc.
Document version	owl:versionInfo	'*n.m*', e.g. '0.1', '1.2'
Date modified	dcterms:modified	Date and time, using W3C profile of ISO 8601 (Wolf & Wicksteed, 1998)
File format	dc:format	Blank node with value of PRONOM Unique Identifier or MIME type, and a label giving name and version as a string
Title	dc:title	String
Author	dc:creator	List of author names as an RDF Sequence or Bag
Approved by	dc:publisher	Name
Date approved	dcterms:dateAccepted	Date and time, using W3C profile of ISO 8601
Access level	dcterms:accessRights	Blank node with value 'Working', '1', '2', etc., and label 'Working draft', 'Public', 'KIM academic members and industrial collaborators', etc.
Keywords	dc:subject	Comma-separated list of keywords
Date created	dcterms:created	Date and time, using W3C profile of ISO 8601
Date issued	dcterms:issued	Date and time, using W3C profile of ISO 8601

for the 'Approved by' property), and the technicalities of how values were handled by available software.

As compared to simply producing documents in the normal way, using the filename register and templates was a little more work but not significantly more, so it was hoped that compliance rates would be high.

Regarding the file naming convention, the central registry was initially implemented as a set of plain text documents, one for each combination of task code and document type, listing the ranked filenames and corresponding titles. This quickly proved unpopular and unwieldy, since registering a new document meant hunting through 135 different files for the right one, then a cycle of downloading, editing, and uploading the file. The register was then reimplemented as a simple web interface which would take the relevant inputs, generate a filename for the user, and log the result. This was much more successful and a high level of compliance was achieved. The only documents that tended to fall outside the file naming convention

and register, as confirmed later when compiling a full list of outputs for the project's final report, were journal submissions.

Regarding the embedded metadata, two sets of Microsoft templates were used: one that enforced the completion of the document properties and one that did not. Some university teams, including the one at the University of Bath, were asked to use the former and the remaining teams used the latter. (The LaTeX templates enforced compliance in that documents would not compile properly if information were missing). The templates that enforced completion were generally successful in collecting the required information, though not necessarily well-liked. With the other templates, a sizeable number of researchers simply added the requested information directly into the document, rather than via the document properties; thus the information was still there, but in a form less amenable to automatic extraction.

One incentive to use the templates as intended was that, if researchers set their file manager to display details of each file, they could add a column to display the document title, as read from the document properties. This would rectify the issue that the file naming convention did not reveal specific details of the content; and since the title could be displayed verbatim was arguably a better solution than having a codified version of the title in the filename. This approach was demonstrated in the display of files in the content management section of the project website, a fact that helped the authors judge levels of compliance. The challenge, however, was that the researchers had to set this up on their own workstations for themselves; the project was not in a position to enforce these settings.

The M3S was therefore a limited success in terms of adoption; it tended to support the team's suspicions that compliance, and quality of compliance, with document and data management protocols would be highest when the effect on the researcher's effort expenditure would be zero or a net decrease. In terms of utility, the planned experiment to extract preservation metadata records from a corpus of project documents did not take place, due to loss of staff towards the end of the project. The central register of filenames did, however, prove enormously helpful in the compilation of the final report and its list of project outputs. When team members were asked for a complete list of their outputs, they had only to fill gaps in an existing list rather than compile a new one from scratch. For those documents conforming to the M3S, the project manager and administrator were able to use information embedded in the filename, along with the title, to judge whether a document should be included and to follow up with the lead author if necessary.

THE RAIDMAP MINIMUM MANDATORY METADATA SET

Five years after the KIM M3S was developed, the authors were engaged on a small, six-month project to design a research data management plan and associated procedures for the Department of Mechanical Engineering at the University of Bath, as an exemplar for similar departments in other institutions (Darlington, 2012). One of the deliverables of this REDm-MED project was RAIDmap (Research Activity Information Development Mapping), a tool to help researchers keep track of their research data outputs. Its primary purpose was to keep a record of the data files produced by a research activity and the associations between them so that, by following the chain of associations, one could see at a glance how a set of published results had been derived from a set of raw data. The motivation was to make it easier to appraise and select data for retention, and to document the data and data processing in such a way as to ensure the reproducibility of the research. A secondary purpose was, once again, to collect enough information about each data file so that, on ingest into a repository, that information could be transformed into standards-compliant preservation and discovery metadata records with minimal effort.

In drawing up the specification for the tool, the question arose of what metadata it should collect about each data file. To answer the question, the authors decided to employ a similar technique to the one they had used in KIM, and to derive an M3S for RAIDmap. It would not have been appropriate to reuse the KIM M3S verbatim as RAIDmap had a quite different set of requirements:

- While the KIM M3S was focused on textual documents, RAIDmap was concerned with data files and databases.
- The KIM approach was intended to be transferable to an industrial context, while RAIDmap was firmly aimed at the academic sector.
- As RAIDmap was intended as a generic tool, not tied to a particular project, it could not rely on a particular file naming convention being used. Neither could a set of project-level common properties be assumed, though a set of common properties might be provided by the user for all files belonging to a given research activity.
- A key part of the RAIDmap proposal was to automate as much of the process of constructing the map as possible. Therefore, some of the metadata extraction that the KIM M3S envisioned happening at the point of ingest into a data archive could and would be performed immediately by the RAIDmap tool.

In addition, in the five years since the KIM M3S was developed, things had moved on in terms of accepted and best practice for preservation and discovery metadata. The DataCite Metadata Schema was achieving acceptance as a cross-discipline

standard for research data discovery metadata (Starr et al., 2011), and PREMIS had reached version 2.1 (PREMIS Editorial Committee, 2011). Therefore the authors took these two schemes as their starting point. Once again they worked through each element in turn, and determined whether it should be supplied by the user, extracted by RAIDmap, or deferred until ingest into a repository, and then whether the information should be recorded against the whole map (called a 'data case' by the team), an individual data record (file or database), or an association between two records (called a 'data development process' by the team).

Derivation of the Set

The analysis of the revised PREMIS scheme resulted in many of the same conclusions as before. The differences regarding elements associated with the PREMIS object entity were as follows:

- **Object Identifier:** RAIDmap would generate one of these for its own internal purposes; with appropriate care, it could be made globally unique.
- **Object Characteristics:**
  - **Format:** RAIDmap would extract this information from the file.
  - **Creating Application:** RAIDmap would extract the name and version of the creating application from the file. Similarly, RAIDmap would determine the date of creation and last modification from either the file itself (if embedded as metadata) or the file system.
  - **Size:** RAIDmap would measure this, but as the archive could also measure it without recourse to the user for correction, the authors decided to treat it as optional.
  - **Inhibitors:** RAIDmap would allow users to record technical restrictions by noting their type, what they restricted, and the password to remove the restriction, though this information would itself be encrypted as a security measure. Since this may not apply, or researchers may prefer to handle this information another way, the team decided not to include it as mandatory, but to support it as an optional element.
- **Storage:** Although in PREMIS this refers to the location of the object after ingest, RAIDmap would need to know the location of the object while in active use.
- **Environment:** The authors considered the creating application to be a sufficient indicator of the environment in most circumstances. In order to cover other cases, the authors proposed two optional elements: software dependency (for specialist plugins, addons, libraries, and so on) and hardware dependency (for specialist hardware). They felt that dependencies between re-

cords would be better recorded in RAIDmap as associations rather than in the record's mandatory metadata.

Regarding the PREMIS event entity, the instances of relevance to the M3S were those represented as data development processes in RAIDmap:

- **Event Type:** This would be recorded in RAIDmap as the name of the data development process.
- **Event Date Time:** The user would be expected to supply this, but RAIDmap would supply a sensible fallback value, such as the creation date and time of a resulting data record.
- **Event Detail:** The user would be able to include this as an annotation on a data development process, but the authors did not consider it mandatory.
- **Linking Agent Identifier:** The user would be expected to supply this, with RAIDmap supplying the user's identity as the fallback value.

The authors recognized that other events relevant to preservation may be recorded elsewhere, such as the project's data management plan.

Regarding the PREMIS agent entity, the authors proposed that the user should be able to maintain an 'address book' of agents within the RAIDmap application so they could be associated with data cases, records and development processes by means of an internal identifier. The properties of the agents were considered out of scope for the M3S, but were expected to include names and contact details.

Regarding the PREMIS rights entity, the authors felt that it would not be efficient to deal with this in full within RAIDmap itself. They proposed instead that researchers provide the details of the rights situation in their data management plan. In cases where different data records had different rights regimes, the authors proposed that each regime be assigned a keyword in the data management plan, and that keyword used as the value of a rights element in RAIDmap for all corresponding data records.

As explained above, PREMIS deferred to descriptive metadata standards for elements relating to intellectual entities, so for RAIDmap the REDm-MED team turned to the DataCite Metadata Schema. Not only was it a suitable generic discovery standard for data in its own right, but the metadata could be used to register a Digital Object Identifier (DOI) for the data record should it be published in an archive.

- **Identifier:** For DataCite this means a DOI. RAIDmap would not be using DOIs, for obvious reasons, but would record an identifier.
- **Creator:** RAIDmap would extract this if possible, using the user's identity as a fallback value.
- **Title:** RAIDmap would extract this if possible.

- **Publisher, Publication Year:** These only apply to published datasets, not to working data records.
- **Subject:** RAIDmap would record this at the levels of project and data case. The idea was that, when the user associated a data case with a project, the subject information for the project would be copied to the data case. The user would then be able to adapt it to suit the particular case.
- **Contributor:** For administrative purposes, the team felt it important that RAIDmap should record information about at least three contributors: first, a contact person for the data record, being the person who works or worked with it most frequently, or has the best understanding of it; second, the data manager responsible for it in a curatorial sense; and third, the person or body who holds the rights to the data. RAIDmap would infer the creator as the most likely contact person. For the other two, the user would supply a default value when setting up a given project and their RAIDmap profile respectively.
- **Date:** RAIDmap would record the dates of creation and last modification, as discussed above.
- **Language:** Depending on the nature of the data this may or may not be relevant, so the team decided this should be optional.
- **Resource Type:** RAIDmap would infer general types such as dataset, image, or model, and some specific types such as spreadsheet.
- **Alternate Identifier:** Presumably the RAIDmap identifier would be demoted to this upon the allocation of an archival or publication identifier by a repository or archive.
- **Related Identifier:** This information would be captured by the associations recorded by RAIDmap, rather than as a property of a data record.
- **Size, Format, Rights:** See the respective discussions above.
- **Version:** The user would supply this information.
- **Description:** The user would supply a brief description of the data record.

This process resulted in four lists of elements. Table 2 shows the mandatory elements for data cases. Table 3 shows the mandatory elements for data records. Table 4 shows the mandatory elements for data development processes. Table 5 shows the optional elements for data records. It will be seen that, having set up the RAIDmap tool with a user profile and project information, in the best case scenario the user would only have to supply a description when adding a data record to a data case. Thus, even though it turned out to be quite a large set for something intended as minimal, the team hoped that completing it would not be burdensome for a researcher.

Due to the tight timescales and limited resources of the project it was not practical or desirable to develop the RAIDmap application entirely from scratch. The team therefore used the Open University's Compendium mind-mapping software

Table 2. Mandatory metadata elements for data cases

Metadata element	Collection method
Project	Supplied by the user from suggestion list
Subject	Initially the subject of associated project

Table 3. Mandatory metadata elements for data records

Metadata element	Collection method
Identifier	Generated by RAIDmap
Title	Determined by RAIDmap, falling back to user entry
Version	Default provided by RAIDmap, corrected if necessary by user
Description	Supplied by the user
Type	Inferred by RAIDmap, corrected if necessary by user
File format (*name and version*)	Determined by RAIDmap, corrected if necessary by user
Creating application (*name and version*)	Determined or inferred by RAIDmap, corrected if necessary by user
Date created	Determined by RAIDmap, corrected if necessary by user
Date modified	Determined by RAIDmap, corrected if necessary by user
Creator	Determined or inferred by RAIDmap, corrected if necessary by user
Contact person	Inferred by RAIDmap, corrected if necessary by user
Data manager	Default provided by RAIDmap, corrected if necessary by user
Rights holder	Default provided by RAIDmap, corrected if necessary by user
Rights	Default provided by RAIDmap, corrected if necessary by user
Filename	Determined by RAIDmap
Location	Determined by RAIDmap, corrected if necessary by user

Table 4. Mandatory metadata elements for data development processes

Metadata element	Collection method
Date and time	Default provided by RAIDmap, corrected if necessary by user
Agent	Default provided by RAIDmap, corrected if necessary by user

Table 5. Optional metadata elements for data records

Metadata element	Collection method
File size	Determined by RAIDmap
Software dependency (*plug-in, add-on*)	Supplied by the user
Hardware dependency	Supplied by the user
Technical restriction (*type and subject of restriction, password*)	Determined by RAIDmap, corrected and extended by the user

as a basis; nodes representing digital records were extended to include a metadata table, populated at creation time by the National Library of New Zealand's Metadata Extractor (Ball & Thangarajah, 2012; Darlington, Thangarajah, & Ball, 2012).

While a working prototype was developed, circumstances did not allow the application to be refined into a production-ready state, so the project team was not able to conduct meaningful tests of its usability and utility. Similarly, code for a RAIDwatch application was written that would monitor the files saved to particular directory tree and automatically add them to a RAID data case in the background. This did not progress to a full working prototype within the project timescale. Nevertheless, trials conducted by the project team indicated that the approach showed some promise.

At the time RAIDmap was developed, the University of Bath did not have an institutional data archive; this came later, in 2015. As of this writing, the focus of the archive is on discovery and bit-level preservation, rather than the active preservation envisioned by PREMIS, therefore the M3S has turned out to be more detailed than would be required for deposit. Having said that, the archive also registers DOIs for datasets, for which purpose the DataCite metadata required by the M3S is of direct relevance.

Comparing the M3S with the metadata collected by the archive, it is notable that the information provided by RAIDmap users would also need to be entered into archival records, either directly or via a 'readme' documentation file. Certain other metadata elements generated by RAIDmap, and the RAID diagram itself, would also make welcome additions to the documentation. While further integration work would be necessary, and a refocusing of the M3S desirable, it is potentially the case that the effort spent by researchers on a RAID diagram would be offset by saved effort at the point of deposit.

RECOMMENDATIONS AND FUTURE RESEARCH DIRECTIONS

Even though the two metadata sets were not thoroughly tested for their intended purpose, they were successful inasmuch as they were implemented with a limited amount of development effort and used with minimal additional human effort. This having been done, the result was a consistent set of information that was amenable to extraction, and that could in principle be transformed into a standardized format relevant to the use cases in question.

The method by which the two sets were constructed may be abstracted thus:

1. Establish a clear use case that the metadata set is seeking to address. With KIM, the driver was the long term preservation of project documents; with RAIDmap, it was the archiving and publication of research data. In both cases, the aim was to collect information at the time of object creation that would make tractable the task of transferring a corpus of such objects to an archive some time later.
2. Identify one or more systems from the wider environment with which to interoperate, and note their metadata requirements. Both the KIM and RAIDmap metadata sets were anticipating the needs of unknown academic sector repositories, and so PREMIS was selected as an appropriate generic target. If the repositories had been known, it would have been better to target their specific requirements. In addition, RAIDmap targeted the DataCite Metadata Schema, being a requirement for obtaining a DOI for a published dataset.
3. Examine the workflow of the user and the environment in which they work, and identify correspondences between the information already in circulation, and the information collected by the target metadata schemes.
4. Identify the most efficient time and place to collect the information. KIM distinguished between information that should be supplied by users when they are actively engaging with the object, by project administrators at the start of the project and at points of change, and by archivists during and after ingest. RAIDmap distinguished between what could be determined automatically, what users should supply once for all their files, and what users should supply about particular files.
5. Find a way to introduce the collection mechanisms into existing workflows and practices such that it either saves the user time and effort, or produces another tangible benefit in exchange for minimal additional effort. The KIM templates were an attempt in this direction, but even the transition to using document properties instead of typing directly into the document seemed to be uncomfortable to the majority of researchers. RAIDmap showed promise for making file organization and data deposit easier, by automatically document-

ing data throughout the active phase of a project. That promise would have been greatly increased had the tool been able to target mature research data management workflows; they were in flux at the time and, while now more settled, they are still evolving.

6. Design application profiles to encode the collected information that both (*a*) make use of and harmonize with metadata structures already in place, and (*b*) match or can be transformed into the format expected by the target metadata schemes. The encoding of the KIM M3S took account of the structures available in common document formats; the encoding of the RAIDmap M3S had more freedom to mimic the requirements and recommendations of PREMIS and DataCite, but still took account of the outputs of the metadata extractor. Had the eventual archives for the outputs been known, it would have been useful to discover their preferred subject terms and suggest them to users as they contributed keywords in the KIM M3S and subjects in the RAIDmap M3S.

The end result is that metadata are recorded in several places with minimal redundancy. This is efficient from the perspective of the Don't Repeat Yourself (DRY) principle: users do not have to duplicate effort providing the same information over and over, and there is less likelihood of inconsistencies being introduced (Hunt & Thomas, 2000). The trade-off is that additional work needs to be performed to assemble a full metadata record from the component sources, and loss of a single source can have a detrimental effect on the whole corpus. This turned out to be an issue in the case of the KIM M3S since, as discussed above, the process of recovering the metadata was not tested, and the matter of encoding the project-level information in a machine-usable form was not addressed.

In hindsight, the KIM work did not go far enough. As well as designing the metadata set and collection mechanisms in tandem, the team should have developed the metadata extraction and transformation mechanisms at the same time, to ensure at the earliest opportunity that the whole workflow would operate as intended. In addition, while some efforts were made to use embedded metadata in project systems, this could have been taken even further to provide greater incentives for compliance.

The avenue explored by RAIDmap warrants further investigation. The prototype tool was able to populate parts of a metadata record automatically for files; there is scope to take that idea further not only for users building an inventory of their own files but to help with uploading that information to repositories. More importantly, the potential of a RAIDwatch-type tool has yet to be fully explored, in particular the ability of a tool to watch changes to a directory and not only add new files to an inventory but also make inferences about the inputs and processes that lead to their creation (McMahon, 2015).

CONCLUSION

Developing a metadata profile for a local application is a task that has many layers. Users must be persuaded to provide the requisite information; the information must be sufficient to support the local application; and it should be possible to transform it such that systems in the wider environment are able to use it as well. The contention of this chapter is that, even if the priorities are in that order, better results may be obtained by tackling the issues in the reverse order. By considering the wider requirements first and then how they compare with local requirements, potential interoperability problems can be avoided and gaps addressed. By allowing the required metadata to be assembled from multiple sources, instead of insisting on a single profile, opportunities open up for highly efficient metadata collection and to reduce the burden on end users. For this to work, though, it is important to design the metadata profiles and supporting systems as a cohesive whole, otherwise the pipeline of metadata from end users to the global infrastructure may break down.

ACKNOWLEDGMENT

The KIM Project was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) and the Economic and Social Research Council (ESRC) under Grant Numbers EP/C534220/1 and RES-331-27-0006. The REDm-MED Project was funded by the Joint Information Systems Committee (JISC) under the Managing Research Data Programme (02) 2011-13.

REFERENCES

Abode Systems. (n.d.). Adobe XMP developer center. Retrieved May 15, 2016, from http://www.adobe.com/devnet/xmp.html

Ball, A., Darlington, M., Howard, T., McMahon, C., & Culley, S. (2012). Visualizing research data records for their better management. *Journal of Digital Information*, *13*(1). Retrieved from https://journals.tdl.org/jodi/article/view/5917/5892

Ball, A., Patel, M., McMahon, C., Green, S., Clarkson, J., & Culley, S. (2006). A grand challenge: Immortal information and through-life knowledge management (KIM). *International Journal of Digital Curation*, *1*(1), 53–59. doi:10.2218/ijdc.v1i1.5

Ball, A., & Thangarajah, U. (2012). *RAIDmap application developer guide*. Retrieved from http://opus.bath.ac.uk/30098

Caplan, P. (2006). Preservation metadata. In S. Ross & M. Day (Eds.), DCC Digital Curation Manual. Edinburgh, UK: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resource/curation-manual/chapters/preservation-metadata/

Consultative Committee for Space Data Systems. (2002). *Reference model for an Open Archival Information System (OAIS)* (Blue Book No. CCSDS 650.0-B-1). Retrieved from http://www.ccsds.org/documents/650x0b1.pdf

Consultative Committee for Space Data Systems. (2012). *Reference model for an Open Archival Information System (OAIS)* (Magenta Book No. CCSDS 650.0-M-2). Retrieved from http://www.ccsds.org/documents/650x0m2.pdf

Darlington, M. (2012). *REDm-MED project final report to JISC*. University of Bath. Retrieved from http://www.ukoln.ac.uk/projects/redm-med/reports/redm-med_final_report_v1.pdf

Darlington, M., Thangarajah, U., & Ball, A. (2012). *RAIDmap application user guide*. University of Bath. Retrieved from http://opus.bath.ac.uk/30097

Day, M. (2004). Preservation metadata. In G. E. Gorman & D. G. Dorner (Eds.), *Metadata applications and management* (pp. 253–273). London: Facet Publishing.

Duranti, L., Eastwood, T., & MacNeil, H. (1997). The preservation of the integrity of electronic records. Vancouver, BC: University of British Columbia; Retrieved from http://www.interpares.org/UBCProject/index.htm

C.U.R.L. Exemplars in Digital Archives. (2000). Metadata for digital preservation: The Cedars Project outline specification. Retrieved from http://www.webarchive.org.uk/wayback/archive/20050111120000/http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html

Heery, R., & Patel, M. (2000). Application profiles: Mixing and matching metadata schemas. *Ariadne*, *25*. Retrieved from http://www.ariadne.ac.uk/issue25/app-profiles/

Hunt, A., & Thomas, D. (2000). *The pragmatic programmer: From journeyman to master*. Reading, MA: Addison-Wesley.

InterPARES project. Project overview. (n. d.). Retrieved from http://www.interpares.org/

Library of Congress. (2016). PREMIS: Preservation metadata maintenance activity. Retrieved from http://www.loc.gov/standards/premis/

Lupovici, C., & Masanès, J. (2000). *Metadata for the long-term preservation of electronic publications*. The Hague, The Netherlands: Koninklijke Bibliotheek. Retrieved from https://www.kb.nl/sites/default/files/docs/NEDLIBmetadata.pdf

McMahon, C. (2015). Design informatics: Supporting engineering design processes with information technology. *Journal of the Indian Institute of Science*, *95*(4), 365–378. Retrieved from http://journal.library.iisc.ernet.in/index.php/iisc/article/view/4585

National Archives of Australia. (1999). Recordkeeping metadata standard for commonwealth agencies. Retrieved from http://pandora.nla.gov.au/nph-wb/20000510130000/http://www.naa.gov.au/www.naa.gov.au/recordkeeping/control/rkms/contents.html

National Archives of Australia. (2015). Australian government recordkeeping metadata standard, version 2.2. Retrieved from http://www.naa.gov.au/records-management/publications/agrkms/

National Library of Australia. (1999). Preservation metadata for digital collections. Retrieved from http://pandora.nla.gov.au/pan/25498/20020625-0000/www.nla.gov.au/preserve/pmeta.html

National Library of New Zealand. (2003). *Metadata standards framework – preservation metadata (revised)*. Retrieved from http://digitalpreservation.natlib.govt.nz/assets/Uploads/nlnz-data-model-final.pdf

OCLC/RLG Preservation Metadata, & the Implementation Strategies Working Group. (2005). *Data dictionary for preservation metadata*. Retrieved from http://www.loc.gov/standards/premis/v1/premis-dd_1.0_2005_May.pdf

OCLC/RLG Working Group on Preservation Metadata. (2001). *Preservation metadata for digital objects: A review of the state of the art*. Retrieved from http://www.oclc.org/content/dam/research/activities/pmwg/presmeta_wp.pdf

OCLC/RLG Working Group on Preservation Metadata. (2002). *Preservation metadata and the oais information model: A metadata framework to support the preservation of digital objects*. Retrieved from http://www.oclc.org/content/dam/research/activities/pmwg/pm_framework.pdf

PREMIS Editorial Committee. (2011). *PREMIS data dictionary for preservation metadata, version 2.1*. Washington, DC: Library of Congress. Retrieved from http://www.loc.gov/standards/premis/v2/premis-2-1.pdf

Public Record Office Victoria. (2000). PROS 99/007 Standard for the management of electronic records, version 1.2. Retrieved from http://pandora.nla.gov.au/pan/22965/20021222-0000/www.prov.vic.gov.au/vers/standards/pros9907.htm

Research Libraries Group, Working Group on Preservation Issues of Metadata. (1998). *Final report*. Retrieved from https://web.archive.org/web/20040216202156/http://www.rlg.org/preserv/presmeta.html

Rogers, C., & Tennis, J. T. (2016). *General Study 15 – Application profile for authenticity metadata*. InterPARES 3 Project. Retrieved from http://www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_gs15_final_report.pdf

Starr, J., Ashton, J., Brase, J., Bracke, P., Gastl, A., Gillet, J., & Ziedorn, F. (2011). *DataCite metadata schema for the publication and citation of research data, version 2.2*. DataCite Consortium; doi:10.5438/0005

University of Pittsburgh, School of Information Sciences. (1996). Metadata specifications derived from the fundamental requirements: A reference model for business acceptable communications. Retrieved from http://web.archive.org/web/20000302194819/www.sis.pitt.edu/~nhprc/meta96.html

Wolf, M., & Wicksteed, C. (1998). *Date and time formats*. World Wide Web Consortium. Retrieved from http://www.w3.org/TR/1998/NOTE-datetime-19980827

KEY TERMS AND DEFINITIONS

Context Information: Information describing how a data object and its Representation Information (q.v.) relate to other information resources.

Data Case: A set of data records associated with some discrete research activity, such as a project, task, or experiment.

Data Development Process: A process that changes or adds to the research data associated with a research activity or project.

Data Record: An object (usually but not always a digital file) that contains data.

Fixity Information: Checksums or other information that could be used to detect, and possibly reverse, undocumented alterations to a data object.

Provenance Information: Information about the source of a data object and its Representation Information (q.v.), the chain of custody since its creation and the operations performed on it.

Reference Information: Unique identifiers for a data object.

Representation Information: The information needed by a user to interpret and understand a data object.

Research Data: Data pertaining to the object of research, in particular those data used as evidence supporting a research conclusion.